Data Analysis Report

Hugo Koopmans

March 29, 2013

Contents

0.1 Introduction

This data analysis report is generated using R, R-studio and knitr to sweave R code and Latex into pdf format. We have the option to include all R code that is used to generate the plots and calculations. Default this feauture is dissabled.
The data analysis step is the first step an a datamining analysis.

0.2 Dataset Basic Artifactes

0.2.1 Basic dataset information

Basic information on dataset:

```
## [1] "/Users/toni/Dropbox/_Scratch/github/dataMineR/1-data-understanding"

## Warning:  cannot open file '1-data-understanding/CopyOfdata-analysis-template.R': No such file
or directory
## Error:  cannot open the connection
```

Read data from file :

Error in eval(expr, envir, enclos) : object 'filename' not found.
The dataset has

Error in eval(expr, envir, enclos) : object 'colums' not found variables and

Error in eval(expr, envir, enclos) : object 'rows' not found rows.
The case identifyer is

Error in eval(expr, envir, enclos) : object 'original_case_id' not found this is unique for all cases.

0.2.2 Excluded variables

From the varables provided the folowing list will be excluded in this anlysis:

Error in eval(expr, envir, enclos) : object 'exclude_var_names' not found

0.2.3 Variabele types

The following variabele are present in the dataset:

Error in eval(expr, envir, enclos) : object 'var_names' not found
Sometimes categoric variables are present as coded numbers. These should be treated as factors. In this dataset the following variables will be used as factors(categoric):

Error in eval(expr, envir, enclos) :
object 'treat_as_categorical' not found
We have

Error in eval(expr, envir, enclos) : object 'num_vars' not found numeric variables and

Error in eval(expr, envir, enclos) : object 'cat_vars' not found categorical variables (or factors in R).

0.3 Numeric variables

Here we analyse all numeric variables. We start with an overview on basic statistics per variable. We check for missing values. We do a histogram plot to show the distribution for this variable. And we test for outliers.

0.3.1 Overview

In the table below we report the number of observations (n), the smallest observation (min), the first quantile (q1), the media , the mean, last quantile, the largest observation (max), and the nber of missing values (na).

```
Error in paste(out, collapse = "\n") :  object 'out' not found
```

0.4 Categorical variables

Here we analyse all categorical variables. We first check the number of different levels in each category(or factor). Then we do a bar plot to show the distribution for each variable.
Overview
In the following table we will see each variable printed with it's unique levels. Beside each level a count is made and a precentage calculated. In the last colum we find a culumative count summing the total up to 100%.

We see that the number of levels can be quite big, for reporting we will omit all variables with more then

```
Error in eval(expr, envir, enclos) :  object 'max_levels' not found
```
levels. These will not be reported in the subsections below.

Variables with to many levels to report are :

```
Error in eval(expr, envir, enclos) :
object 'cat_var_names_not_reported' not found
```

```
Error in paste(out, collapse = "\n") :  object 'out' not found
```

0.5 Behavioural Analysis

The next step is behavioural analysis. The current dataset is now saved.

Dataset saved as :

```
Error in eval(expr, envir, enclos) :  object 'datasetName' not found
```